PSTAT 174 Final Project: Industrial Production Index of Electric and Gas Utilities

Kristian Abad

12/1/2021

Abstract

The Industrial Production (IP) index in this case is a measure of energy production from electric and gas utilities in which questions of whether or not there is seasonality in the time series as well as whether we can predict energy production. The techniques used include log transformation, differencing at 2 different lags in order to remove trend and seasonality, and model selection based off AICC. All in all, more work would be needed to get a better fitting model.

Introduction

The data set used holds industrial production of electric and gas utilities in the United States from 1985-2018 containing monthly data obtained via Kaggle. The problem here is testing whether or not data can be forecasted. Despite log transformations, differencing, and selection of a fit model based off AICC, diagnostics proved to be unsatisfactory which was key in determining that more work and information would be needed to make the model better as would show in the forecast stage. The model proved to be somewhat accurate yet underestimated the true values.

```
library(dplyr)
```

Plot and analysis of the time series

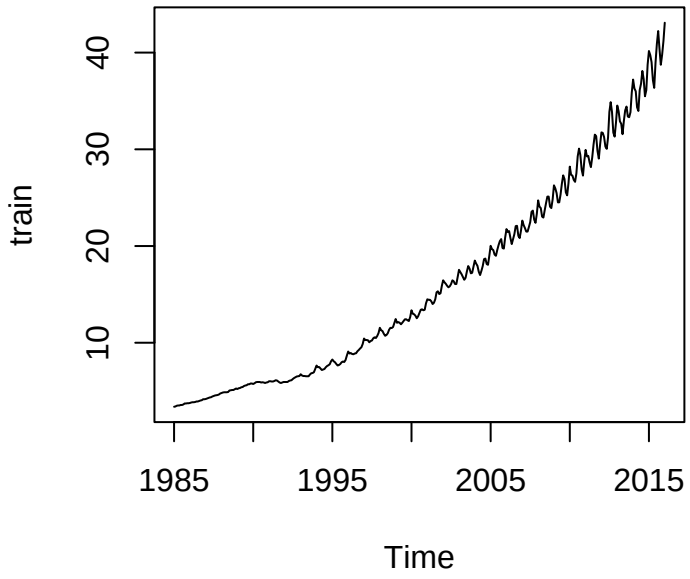
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
setwd("D:/Users/Kris 2/Documents/Fall 2021/PSTAT 174/Final Course Project")
industrial_Prod <- read.delim("IPG2211A2N.csv", header = TRUE, sep = ",") %>% rename(date = DATE, Energy
#head(industrial_Prod)
```

```r
#Suppose we only need the values
industrial_Prod_Mod <- industrial_Prod$Energy_Production
head(industrial_Prod_Mod)
```

```
## [1] 3.3842 3.4100 3.4875 3.5133 3.5133 3.5650
```

```r
ts_IndustrialProduction <- ts(data = industrial_Prod_Mod,
                              start = c(1985,1),
                              end = 2018,
                              frequency = 12)
#plot(industrial_Prod$date,industrial_Prod$Energy_Production,xlab = "Year",ylab="US population (in Mill
train <- ts(data = ts_IndustrialProduction[1:384],
            start = c(1985,1),
            end = 2016,
            frequency = 12)
#The one obs in 2018 wasn't included for consistency
test <- ts(data = ts_IndustrialProduction[385:396],
            start = c(2017,1),
            frequency = 12)
plot(train)
```



(i) Based off the plot, there does appear to be a positive trend as the year increases.
(ii) There does appear to be a little bit of a seasonal component.
(iii) There aren't any apparent sharp changes in behavior that is worth mentioning. However the trend makes the time series look like it has some sort of quadratic relationship due to the parabola.

Model identification

```r
#Transformations to stabilize variance/seasonality effects
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
cat("Variance before any transform: ", var(train))
```

```
## Variance before any transform:  113.1331
```

```r
cat("\nVariance after log transform: ", var(log(train)))
```

```
##
## Variance after log transform:  0.5180023
```

```r
t <- 1:length(train)
bcTransform = boxcox(train ~ t,plotit = FALSE)
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
data.bc = (1/lambda)*(train^lambda-1)
cat("\nVariance after box-cox transform: ", var(data.bc))
```

```
##
## Variance after box-cox transform:  0.8580174
```

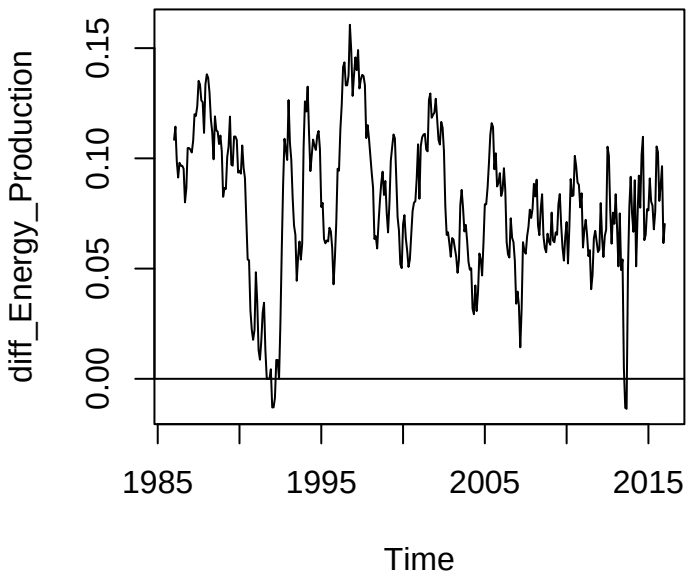
```r
train <- log(train)
diff_Energy_Production <- diff(train,12)
cat("\nVariance after diff. at lag 12: ", var(diff_Energy_Production))
```

```
##
## Variance after diff. at lag 12:  0.001053672
```

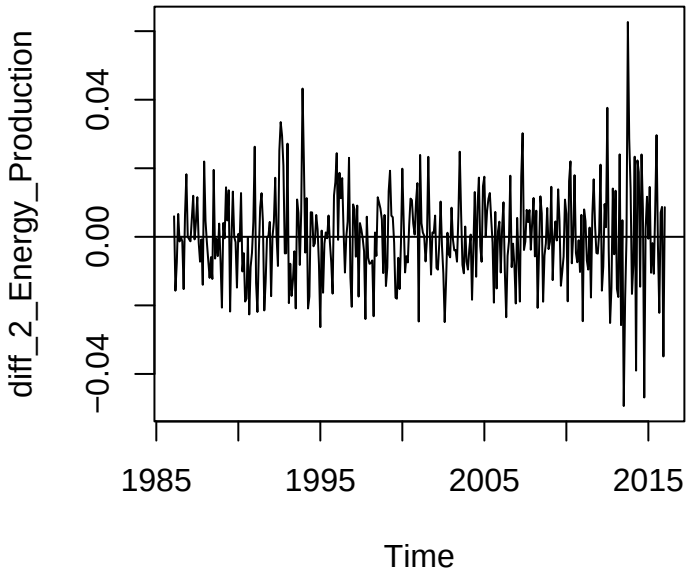
```r
diff_2_Energy_Production <- diff(diff_Energy_Production,1)
cat("\nVariance after diff. at lag 1: ", var(diff_2_Energy_Production))
```

```
##
## Variance after diff. at lag 1:  0.0001872279
```

```r
ts.plot(diff_Energy_Production)
abline(h=0)
```



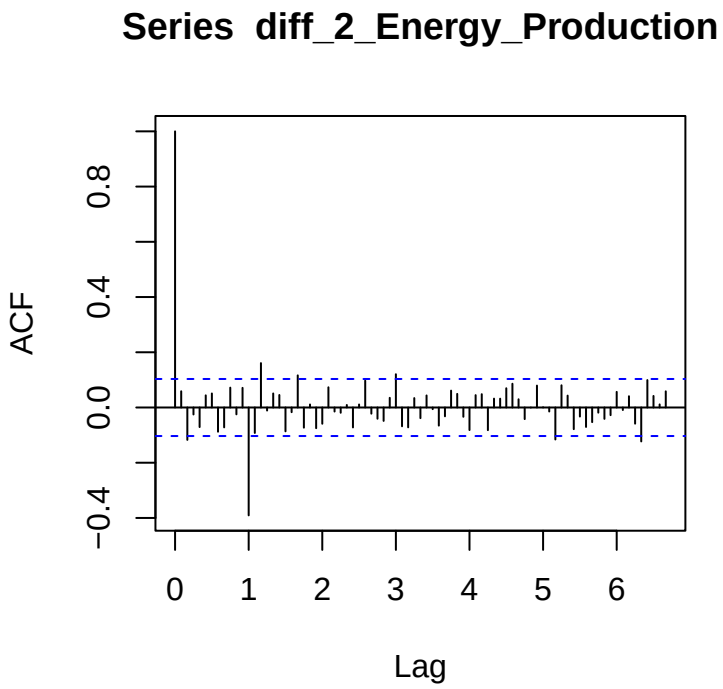

```
ts.plot(diff_2_Energy_Production)
abline(h=0)
```



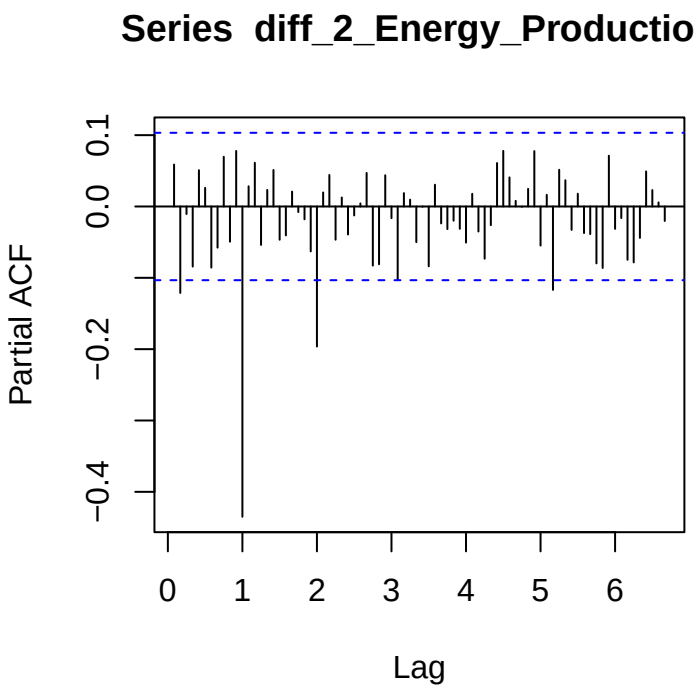
Differencing at lag 1 was used in order to remove the positive trend. There did appear to be some seasonality and so differencing at lag 12 was utilized since the frequency of the data was monthly. The time series does

look to be stationary after differencing was applied. Additionally, variance was tracked along any change to the time series in order to determine if it was better to work with the original data and to ensure there was no overdifferencing.

```
acf(diff_2_Energy_Production,lag.max = 80)
```




```
pacf(diff_2_Energy_Production, lag.max = 80)
```



The ACF graph indicates an order of one because at lag 1 we have a nonzero value and for the rest of the remaining integers the are within the confidence band or are zero. For the PACF graph, the nonzero values occur at lags one and come pretty close to 5.The possible models are SARIMA(2,1,1)x(0,1,0), SARIMA(5,1,1)x(0,1,0).

Fitting the model

```
library(qpcR)
```

```
## Warning: package 'qpcR' was built under R version 4.1.2
```

```
## Loading required package: minpack.lm
```

```
## Warning: package 'minpack.lm' was built under R version 4.1.2
```

```
## Loading required package: rgl
```

```
## Warning: package 'rgl' was built under R version 4.1.2
```

```
## Loading required package: robustbase
```

```
## Warning: package 'robustbase' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
cat("AR ","MA", "\n")
```

```
## AR  MA
```

```
for (i in 0:2){
  for (j in 0:1){

    # print(i);
    # print(j);
    aicc <- AICc(arima(diff_2_Energy_Production,
                       order = c(i,1,j),
                       seasonal = list(order = c(0,1,0),period = 12),
                       method = "ML"))

    cat(i, " ",j," ", aicc, "\n")
  }
}
```

```
## 0   0   -1412.368
## 0   1   -1625.083
## 1   0   -1471.217
## 1   1   -1624.244
## 2   0   -1519.265
## 2   1   -1628.378
```

```
library(qpcR)
# x.ywest=ar(diff_2_Energy_Production, aic = TRUE, order.max = NULL, method = c("yule-walker"))
# x.ywest
# sqrt(diag(x.ywest$asy.var.coef))
model <- arima(diff_2_Energy_Production,
               order=c(1,1,0),
               seasonal = list(order = c(0,1,0),period = 12),
               #fixed = c(NA,NA,NA), #here you can set coeff to zero
               method="ML",)
model
```

```
##
## Call:
## arima(x = diff_2_Energy_Production, order = c(1, 1, 0), seasonal = list(order = c(0,
##     1, 0), period = 12), method = "ML")
##
## Coefficients:
##           ar1
##       -0.4027
## s.e.   0.0494
##
## sigma^2 estimated as 0.0008336:  log likelihood = 737.61,  aic = -1471.23
```

```
model2 <- arima(diff_2_Energy_Production,
                order=c(2,1,0),
                seasonal = list(order = c(0,1,0), period = 12),
                fixed = c(NA,NA),
                method="ML")
model2
```

```
##
## Call:
## arima(x = diff_2_Energy_Production, order = c(2, 1, 0), seasonal = list(order = c(0,
##     1, 0), period = 12), fixed = c(NA, NA), method = "ML")
##
## Coefficients:
##           ar1      ar2
##       -0.5486  -0.3682
## s.e.   0.0501   0.0501
##
## sigma^2 estimated as 0.000721:  log likelihood = 762.65,  aic = -1519.3
```

```
AICc(model)
```

```
## [1] -1471.217
```

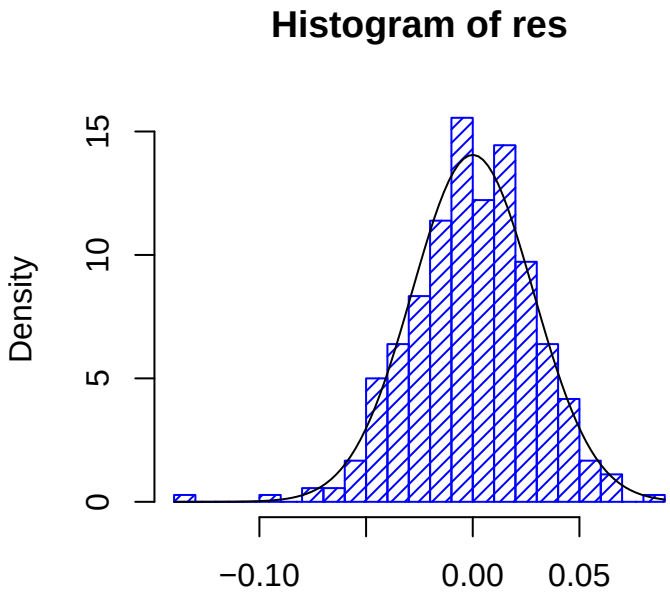
```
AICc(model2)
```

```
## [1] -1519.265
```

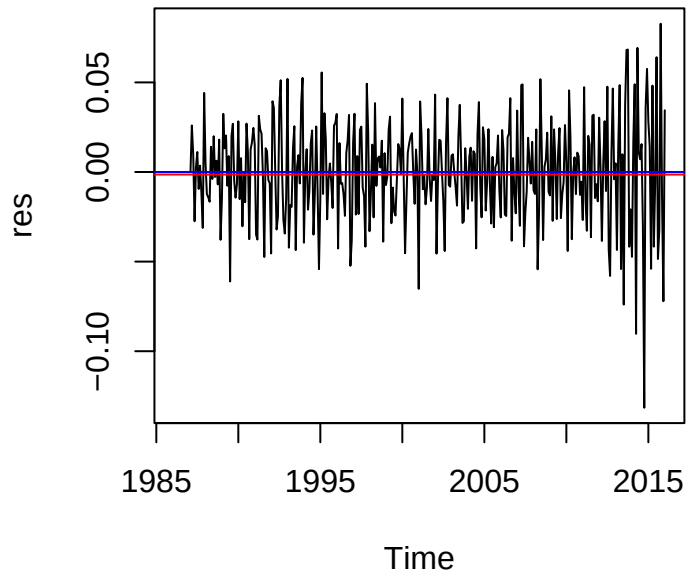
According to the AICC values, model 1 or SARIMA(1,1,0)x(0,1,0) is the "best" model which differs from what was initially thought to be the model. However, we now move on to diagnostic checks.

Diagnostic checks for the first model (SARIMA(1,1,0)x(0,1,0)):

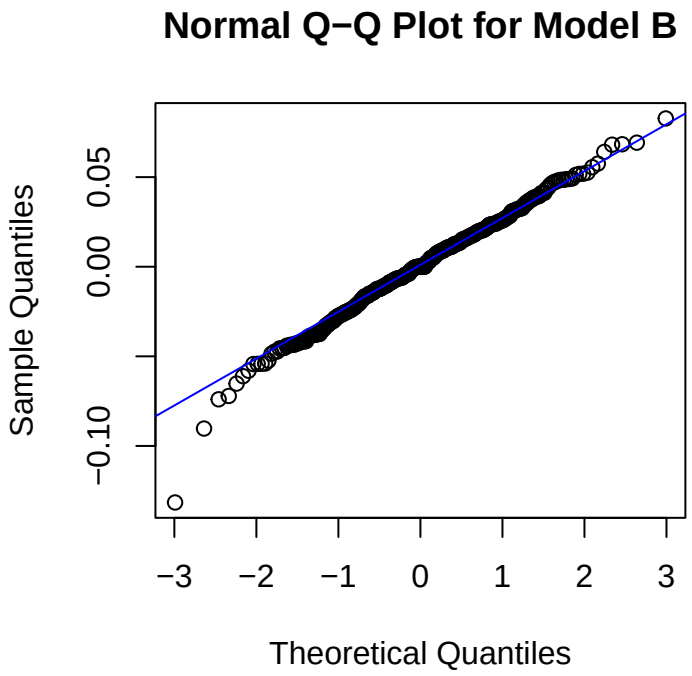
```
res <- residuals(model)
hist(res,density=20,breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res)
std <- sqrt(var(res))
curve( dnorm(x,m,std), add=TRUE )
```



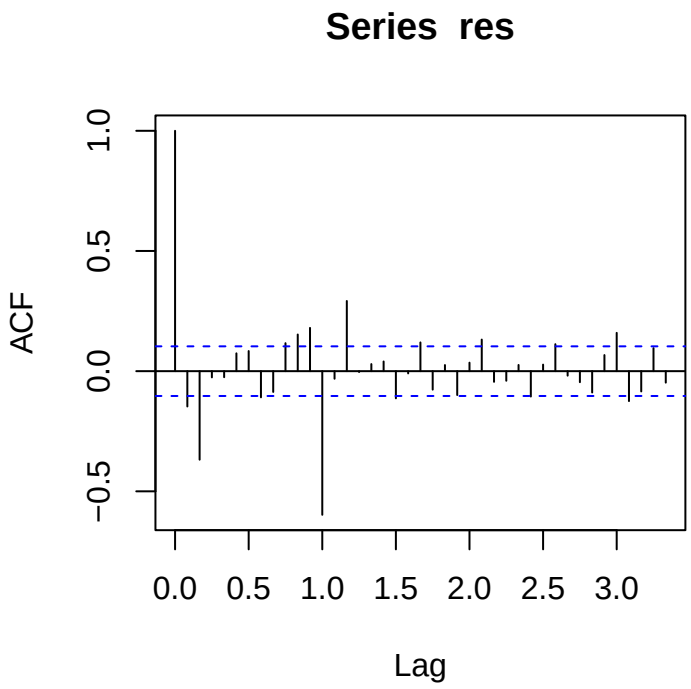

```
plot.ts(res)
fitt <- lm(res ~ as.numeric(1:length(res))); abline(fitt, col="red")
abline(h=mean(res), col="blue")
```



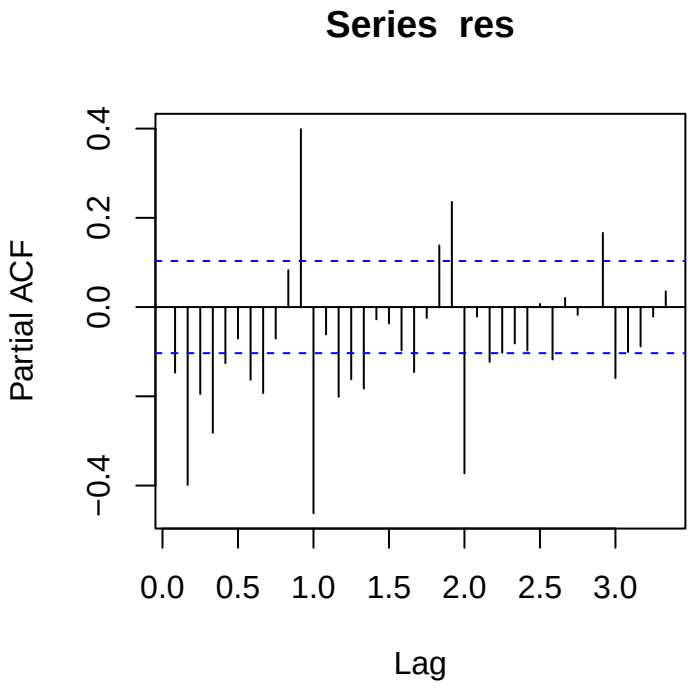
```
qqnorm(res,main= "Normal Q-Q Plot for Model B")
qqline(res,col="blue")
```




```
acf(res, lag.max=40)
```




```
pacf(res, lag.max=40)
```



```r
Box.test(res, lag = 19, type = c("Box-Pierce"), fitdf = 1)
```

```
##
##  Box-Pierce test
##
## data:  res
## X-squared = 258.74, df = 18, p-value < 2.2e-16
```

```r
Box.test(res, lag = 19, type = c("Ljung-Box"), fitdf = 1)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 267.44, df = 18, p-value < 2.2e-16
```

```r
Box.test(res^2, lag = 19, type = c("Ljung-Box"), fitdf = 1)
```

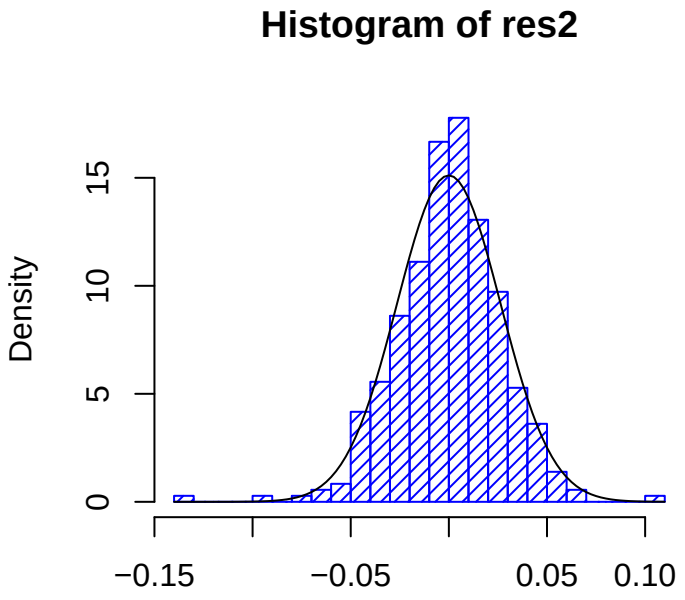
```
##
##  Box-Ljung test
##
## data:  res^2
## X-squared = 228.44, df = 18, p-value < 2.2e-16
```

```r
#acf(res^2, lag.max=40)
```

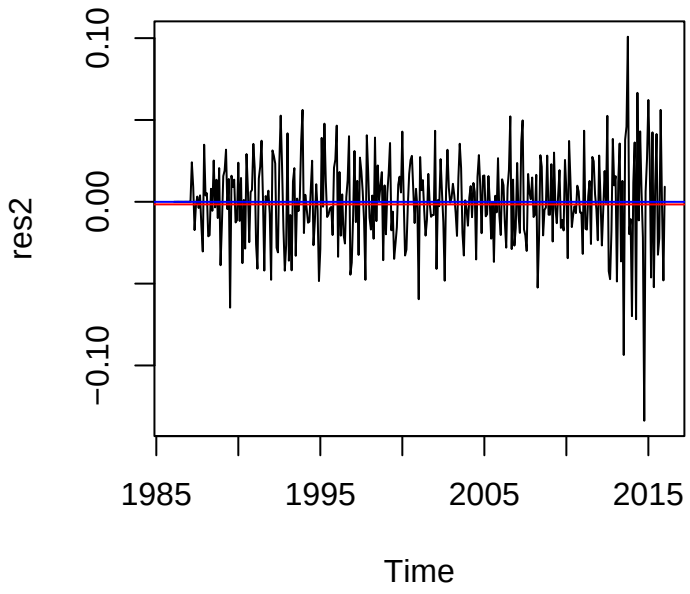
The plot of the residuals doesn't any show seasonality, normal qq plot fails the normal assumption, and the acf/pacf graphs show the residuals to be nonzero, and the p-values are greater than 0.05 and thus failing the diagnostic checks in multiple ways.

Diagnostic checks for model 2:

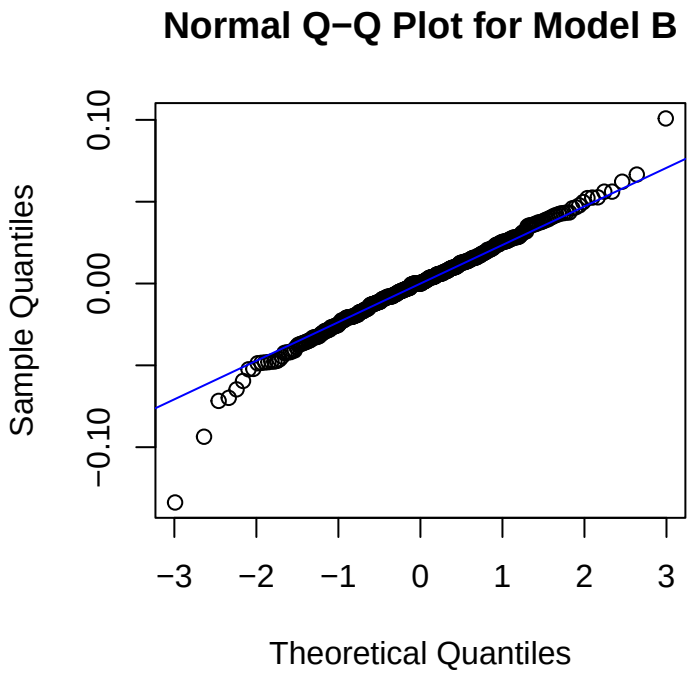
```r
res2 <- residuals(model2)
hist(res2,density=20,breaks=20, col="blue", xlab="", prob=TRUE)
m2 <- mean(res2)
std2 <- sqrt(var(res2))
curve( dnorm(x,m2,std2), add=TRUE )
```



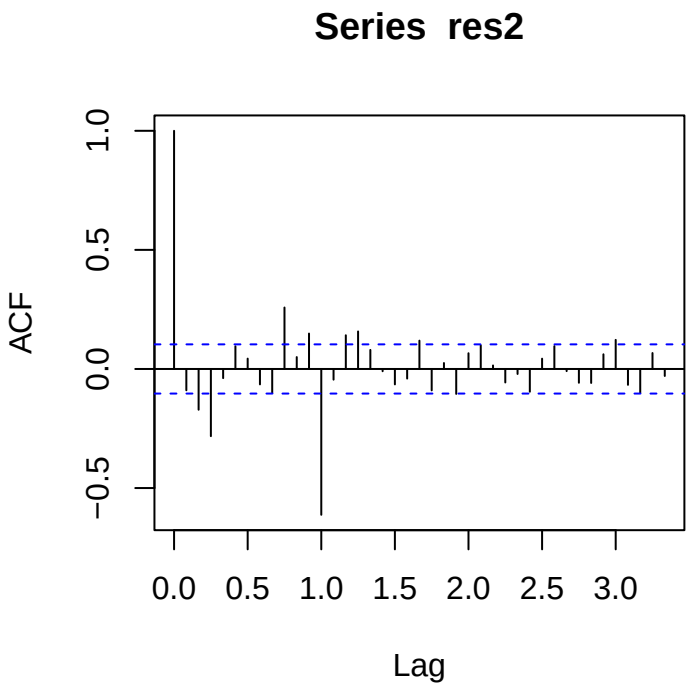

```
plot.ts(res2)
fitt <- lm(res2 ~ as.numeric(1:length(res2))); abline(fitt, col="red")
abline(h=mean(res2), col="blue")
```



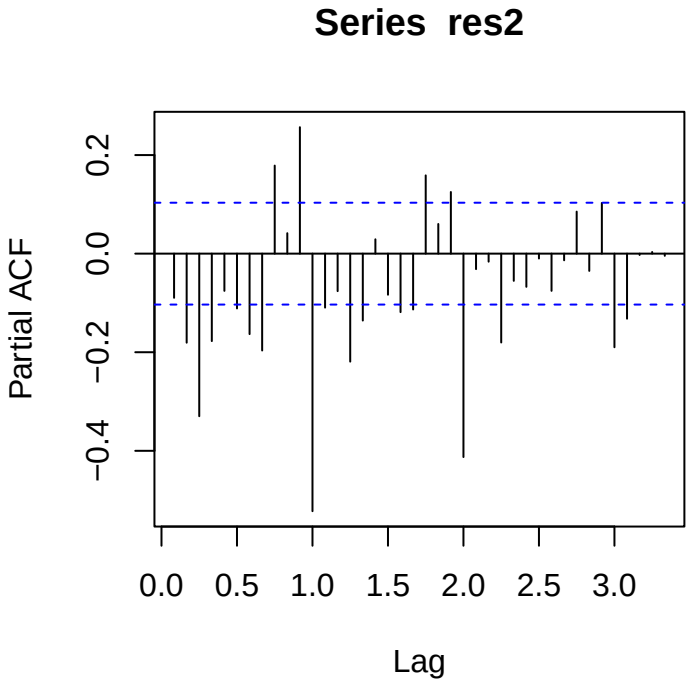
```r
qqnorm(res2,main= "Normal Q-Q Plot for Model B")
qqline(res2,col="blue")
```




```r
acf(res2, lag.max=40)
```




```
pacf(res2, lag.max=40)
```



```
Box.test(res, lag = 19, type = c("Box-Pierce"), fitdf = 2)
```

```
##
##  Box-Pierce test
##
## data:  res
## X-squared = 258.74, df = 17, p-value < 2.2e-16
```

```
Box.test(res, lag = 19, type = c("Ljung-Box"), fitdf = 2)
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 267.44, df = 17, p-value < 2.2e-16
```

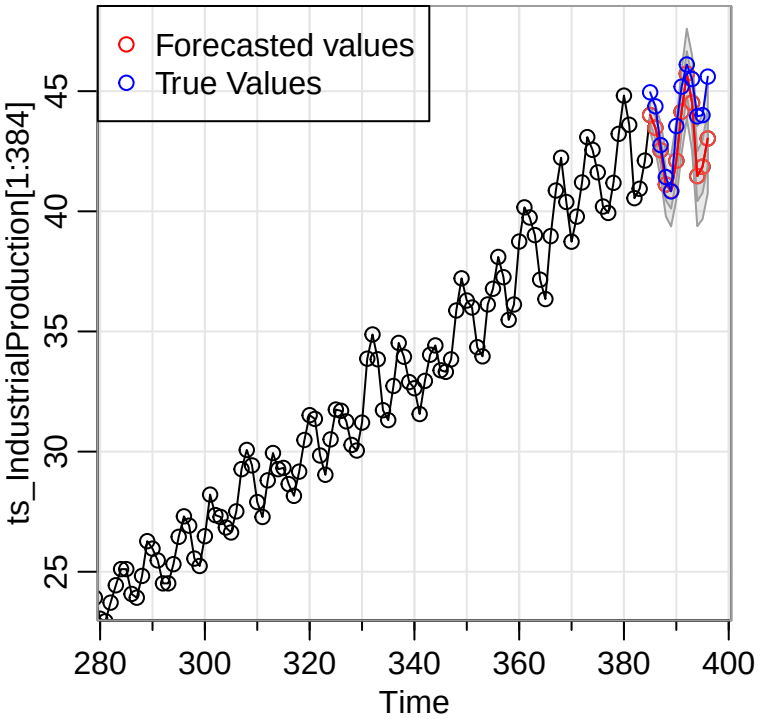
```
Box.test(res^2, lag = 19, type = c("Ljung-Box"), fitdf = 2)
```

```
##
##  Box-Ljung test
##
## data:  res^2
## X-squared = 228.44, df = 17, p-value < 2.2e-16
```

The analysis of the residuals proved to be unsatisfactory.This model still failed diagnostics similarly to the first model. Despite the log-transforms and differencing, nothing seemed to improve the diagnostic checks even setting some coefficients to zero. The "model" used for forecasting in this case is the first model due to low AICC which is $(1 - 0.4027B)X_t = Z_t$.

Forecasting

```
library(astsa)
pred.tr <- sarima.for(ts_IndustrialProduction[1:384], n.ahead=12, plot.all=F,
p=1, d=1, q=0, P=0, D=1, Q=0, S=12)
lines(385:396, as.vector(pred.tr$pred), col="red")
lines(385:396, ts_IndustrialProduction[385:396], col="blue")
points(385:396, ts_IndustrialProduction[385:396], col="blue")
legend("topleft", pch=1, col=c("red", "blue"),
legend=c("Forecasted values", "True Values"))
```



Conclusion

In conclusion, the time series appeared to show some seasonality. However, when it came to forecasting, the model chosen could be improved as based off the forecasts, the model underestimates the actual values. This could most likely improve my tweaking the model as lots of methods were already exhausted to rectify the diagnostics. More information and methods would be needed to improve the model.

References

https://www.kaggle.com/sadeght/industrial-production-electric-and-gas-utilities